Ultrahigh Dimensional Feature Screening via RKHS Embeddings

Krishnakumar Balasubramanian
Georgia Institute of Technology
krishnakumar3@gatech.edu

Bharath K. Sriperumbudur
University of Cambridge
bs493@statslab.cam.ac.uk

Guy Lebanon
Georgia Institute of Technology
lebanon@cc.gatech.edu

Abstract

Feature screening is a key step in handling ultrahigh dimensional data sets that are ubiquitous in modern statistical problems. Over the last decade, convex relaxation based approaches (e.g., Lasso/sparse additive model) have been extensively developed and analyzed for feature selection in high dimensional regime. But in the ultrahigh dimensional regime, these approaches suffer from several problems, both computationally and statistically. To overcome these issues, in this paper, we propose a novel Hilbert space embedding based approach to independence screening for ultrahigh dimensional data sets. The proposed approach is *model-free* (i.e., no model assumption is made between response and predictors) and could handle non-standard (e.g., graphs) and multivariate outputs directly. We establish the sure screening property of the proposed approach in the ultrahigh dimensional regime, and experimentally demonstrate its advantages and superiority over other approaches on several synthetic and real data sets.

1 Introduction

Ultrahigh dimensional data sets are ubiquitous in modern statistical problems arising from several diverse scientific fields. For example, several biological problems or high frequency trading problems have several million features (denoted as p) compared to a much lesser number of samples (denoted as n). Feature screening plays an important role in analyzing these 'large p small n' data sets. Various penalization based techniques that promote sparsity have been developed and analyzed in this regime: Lasso (Tibshirani, 1996), Dantzig selector (Candes and Tao, 2007) and scad penalties (Fan and Li, 2001) assume a linear model between the covariates and the response, while SPAM and related techniques (Ravikumar et al., 2009, Huang et al., 2010) assume a non-linear model in order to select a few relevant features. All these methods allow for the data dimensionality to be greater than the sample size.

However, there are several issues with the above mentioned penalty approaches in ultrahigh dimensions. First, these methods cannot efficiently handle ultrahigh dimensional settings with p growing faster than a polynomial rate in n, e.g., p growing exponential in n. Secondly, the irrepresentability conditions (Zhao and Yu, 2007)—these conditions mean that the covariates not in the true model are not representable, in some sense, by the covariates in the true model—under which the model selection consistency is proved for the penalty methods in high-dimensions, are too stringent to hold in ultrahigh dimensions (see Fan and Lv, 2010, Section 5.5 for general discussion about this and concrete examples). Thirdly, penalization approaches are computationally expensive, e.g., a typical algorithm for lasso scales as $O(p^3)$ with other parameters fixed, hence expensive for ultrahigh dimensional p problems.

In order to tackle this situation, an alternate line of research based on marginal regression was proposed and analyzed (Fan and Lv, 2008, Fan et al., 2009). This is a relatively old technique, that has re-emerged as an alternative for feature screening in ultrahigh dimensions. The general idea of this approach is to measure the relationship (to be clearly defined based on context) of each feature individually to the response and rank them accordingly. For example, assuming a linear model between response and covariates, Fan and Lv (2008) proposed to measure the residual between response and each covariate (in a least-square sense) and rank the covariates accordingly. In order to relax the linear model assumption, Fan et al. (2009) proposed screening for generalized linear models based on marginal utility; Fan et al. (2011) proposed screening using a non-parametric additive model based on

Appearing in Proceedings of the 16^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2013, Scottsdale, AZ, USA. Volume 31 of JMLR: W&CP 31.

smoothing splines. Recently, Li et al. (2012) proposed a model-free (i.e., without any regressive modeling assumptions) screening procedure, DC-SIS, based on distance covariance metric (Székely et al., 2007)—which is zero if and only if the random variables are independent—as a measure of relationship between response and covariate. To elaborate, if the distance covariance between the response and a covariate is "small", then the response is independent of the covariate and therefore such a covariate can be screened out from consideration. Recently Ji and Jin (2012) showed that a two-step procedure—screening followed by penalized regression—is optimal for feature selection in this regime.

In this paper, we propose a general framework, *sup*-HSIC-SIS (Hilbert Schmidt independence criterion–Sure independence screening), for model-free, multi-output screening. The approach uses RKHS based independence measures (Gretton et al., 2005) and generalizes the previously proposed DC-SIS approach. This proposal is motivated from the recent work by Sejdinovic et al. (2012a) who established the equivalence between distance covariance and HSIC (a dependence/independence measure based on RKHS embedding of probabilities). Given this equivalence, it is straight forward to propose an independence screening procedure based on HSIC by replacing distance covariance in (Li et al., 2012) with HSIC and carrying out the analysis verbatim. However, a major issue with DC-SIS (or its equivalent RKHS version, say HSIC-SIS) is that the employed independence measure is just one member of a parametric family of independence measures and there is no guarantee that this member provides the best screening procedure over all the other choices from this family. In other words, if we consider HSIC-SIS, the choice of kernel determines the performance of the screening procedure.

Our main contribution in this paper is to address this issue by using an independence measure (that adapts to the joint distribution between the response and covariates) that is obtained by taking the supremum of HSIC over a family of kernels, and theoretically show that *sup*-HSIC-SIS enjoys the sure screening property under some regularity conditions. Furthermore, we propose two iterative versions of *sup*-HSIC-SIS that address issues inherent in any marginal screening procedure and are robust to the assumed regularity conditions. We empirically show that these proposed extensions along with *sup*-HSIC-SIS perform better than the existing approaches, while the theoretical analysis of these extensions are left out for future work.

A related RKHS based approach was previously proposed for feature selection in Song et al. (2012). The approach uses HSIC metric and deals primarily with the low-dimensional setting (i.e., $n > p$) and is basically a model-free version of subset selection approaches used in linear regression settings. Comparing their empirical results with ours (see Sections 6.4 and 6.5), we note that while BA-HSIC is suitable for low-dimensions and to some extent for high-dimensional settings, it does not perform well in ultrahigh dimensional settings. In addition, while (Song et al., 2012) do not provide any theoretical guarantees for their approach, we conjecture that BA-HSIC performs inferior to DC-SIS and *sup*-HSIC-SIS in ultrahigh dimensional settings using the arguments similar to the ones used in (Li et al., 2012).

The paper is organized as follows. In Section 2, we introduce the *sup*-HSIC dependence measure. In Section 3, we discuss how it could be used for feature screening in ultra-high dimensions. The sure screening property of *sup*-HSIC-SIS is then analyzed in Section 4 and two related iterative extensions are discussed in Section 5. Experimental results comparing the proposed methods with various other approaches on synthetic and real-world data sets are provided in Section 6. Missing proofs are provided in an appendix.

2 RKHS embedding of probabilities

Recently, the notion of embedding probability measures into a reproducing kernel Hilbert space (RKHS) has been proposed as a generalization to the classical kernel method (which embeds points from an input space into an RKHS) with a motivation to provide a linear method for handling higher-order statistics of random variables (Berlinet and Thomas-Agnan, 2004, Smola et al., 2007). This notion has gained popularity in various applications including hypothesis testing, dimensionality reduction and reinforcement learning (see Nishiyama et al., 2012, and references therein). Formally, given a Borel probability measure, $\mathbb{P}$ defined on a topological space, $\mathcal{X}$, and the RKHS $(\mathcal{H}, k)$ of functions on $\mathcal{X}$ with bounded and measurable k as its reproducing kernel, the embedding of $\mathbb{P}$ into $\mathcal{H}$ is defined as $\mathbb{P}k := \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x)$. Given two Borel probability measures, $\mathbb{P}$ and $\mathbb{Q}$, Gretton et al. (2007) defined the RKHS distance between their embeddings as the *maximum mean discrepancy* (MMD), i.e.,

$$\gamma_k(\mathbb{P}, \mathbb{Q}) \stackrel{\text{def}}{=} \left\| \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x) - \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{Q}(x) \right\|_{\mathcal{H}}.$$

Note that when the kernel k is characteristic (Sriperumbudur et al., 2010), the embeddings are injective, i.e., $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0$ if and only if $\mathbb{P} = \mathbb{Q}$ and thus γ_k defines a metric on the space of probability measures. One of the applications of the above metric is in capturing the degree of dependence between two random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ with marginal distributions $\mathbb{P}^X$ and $\mathbb{P}^Y$ and jointly distributed as $\mathbb{P}^{XY}$. Assuming $k : (\mathcal{X} \times \mathcal{Y})^2 \to \mathbb{R}$ to be separa-

ble, i.e., $k((x,y),(x',y')) = k_{\mathcal{X}}(x,x')k_{\mathcal{Y}}(y,y')$, where $k_{\mathcal{X}} : \mathcal{X}^2 \to \mathbb{R}$ and $k_{\mathcal{Y}} : \mathcal{Y}^2 \to \mathbb{R}$ are reproducing kernels of $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$ respectively (so that $\mathcal{H} \cong \mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$), γ_k^2 reduces to the Hilbert-Schmidt independence criterion (Gretton et al., 2005) between X and Y, defined as

$$\begin{aligned}\gamma_k^2(\mathbb{P}^{XY}, \mathbb{P}^X\mathbb{P}^Y) &\stackrel{\text{def}}{=} \|\mathbb{P}^{XY}k - \mathbb{P}^X\mathbb{P}^Y k\|_{\mathcal{H}}^2 \\ &= \mathsf{E}_{XX'YY'}\left[k_{\mathcal{X}}(X,X')k_{\mathcal{Y}}(Y,Y')\right] \\ &+ \mathsf{E}_{XX'}\left[k_{\mathcal{X}}(X,X')\right]\mathsf{E}_{YY'}\left[k_{\mathcal{Y}}(Y,Y')\right] \\ &- 2\mathsf{E}_{XY}\left[\mathsf{E}_{X'}[k_{\mathcal{X}}(X,X')]\mathsf{E}_{Y'}[k_{\mathcal{Y}}(Y,Y')]\right], \end{aligned} \quad (1)$$

where X' and Y' are independent copies of X and Y respectively. Gretton et al. (2005) showed that $\gamma_k(\mathbb{P}^{XY}, \mathbb{P}^X\mathbb{P}^Y)$ is the Hilbert-Schmidt norm of the cross-covariance operator between $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$, with the property that when $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ are characteristic: $\gamma_k(\mathbb{P}^{XY}, \mathbb{P}^X\mathbb{P}^Y)$ is zero iff X and Y are independent. This crucial property of γ_k will be exploited later in our screening framework. A drawback of the above metric is: typically the kernel comes with a tuning parameter that should be selected in practice using heuristics. In order to deal with this problem, Sriperumbudur et al. (2009) proposed the following modification (actually proposed in the context of MMD, which we here present in the context of HSIC) which we call as *sup*-HSIC:

$$\gamma(\mathbb{P}^{XY}, \mathbb{P}^X\mathbb{P}^Y) \stackrel{\text{def}}{=} \sup\{\gamma_k(\mathbb{P}^{XY}, \mathbb{P}^X\mathbb{P}^Y) : k \in \mathcal{K}\}.$$

Note that γ represents the maximal distance between $\mathbb{P}^{XY}$ and $\mathbb{P}^X\mathbb{P}^Y$ over the family of kernels $\mathcal{K}$. If any $k \in \mathcal{K}$ is characteristic, then γ is a metric. Typical example includes the family of Gaussian kernels $\mathcal{K}_G(u,v)$ when $k_{\mathcal{X}} = k_{\mathcal{Y}} \stackrel{\text{def}}{=} \{\exp^{-\sigma\|u-v\|_2^2} : \sigma \in \mathbb{R}_+\}$. See (Sriperumbudur et al., 2009) for more details and examples.

In statistical problems, we are given n random samples $\{(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\}$ drawn i.i.d. from $\mathbb{P}^{XY}$. Given these samples, an estimate $\widehat{\gamma}$ of *sup*-HSIC is defined as:

$$\begin{aligned}\widehat{\gamma}(\mathbb{P}^{XY}, \mathbb{P}^X\mathbb{P}^Y) &\stackrel{\text{def}}{=} \sup\{\|\mathbb{P}_n^{XY}k - \mathbb{P}_n^X\mathbb{P}_n^Y k\|_{\mathcal{H}} : k \in \mathcal{K}\}, \\ &= \frac{1}{n}\sup_{k_{\mathcal{X}}\in\mathcal{K}_{\mathcal{X}}, k_{\mathcal{Y}}\in\mathcal{K}_{\mathcal{Y}}} \sqrt{\operatorname{trace}\left(\mathbf{K}_{\mathcal{X}}\mathbf{H}\ \mathbf{K}_{\mathcal{Y}}\mathbf{H}\right)}\end{aligned}$$

where $\mathbb{P}_n^{XY}, \mathbb{P}_n^X$ and $\mathbb{P}_n^Y$ represent the empirical measures over the given samples, $\mathbf{K}_{\mathcal{X}}$ and $\mathbf{K}_{\mathcal{Y}}$ are $n \times n$ Gram matrices associated with $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ respectively, $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ where $\mathbf{I}$ is $n \times n$ identity matrix and $\mathbf{1}$ is a $n \times 1$ vector of ones.

3 Screening via RKHS embedding

In this section, we describe how the *sup*-HSIC measure of independence could be used for feature screening in ultrahigh dimensions. We assume a response $Y \in \mathbb{R}^d$ and covariates $X \in \mathbb{R}^{p_n}$, with p_n growing with n and d fixed (for simplicity). The method applies as well to more general topological spaces $\mathcal{X}, \mathcal{Y}$. We use X_r to denote the r-component of X and $X_{\mathcal{S}}$ to denote the components of X indexed by the elements of the set $\mathcal{S}$. We denote the n training set samples as $\{(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\}$ where n can be very small compared to p_n. Under such an assumption, it is natural to assume that only a subset of covariates are related to the response Y.

Following Li et al. (2012), we define the set of relevant variables $\mathcal{M}$ and irrelevant variables $\mathcal{I}$ as:

$$\begin{aligned}\mathcal{M} &= \{r : \mathbb{P}(Y|X) \text{ depends on } X_r\} \\ \mathcal{I} &= \{r : \mathbb{P}(Y|X) \text{ does not depend on } X_r\}\end{aligned}$$

where $\mathbb{P}(Y|X)$ is the conditional distribution of Y given X. Note that given $X_{\mathcal{M}}$, $X_{\mathcal{I}}$ is conditionally independent of Y and hence redundant while calculating the response. With this definition, feature selection involves estimating the set $\mathcal{M}$ from the given n samples.

A natural idea is to rank the covariates according to their degree of dependence to the response. In order to measure such a degree of dependence of the dimension X_r to Y, we use the *sup*-HSIC measure introduced in the previous section. Specifically, we use the *sup*-HSIC between the joint random variable (X_r, Y) and the marginals X_r and Y. Denoting the joint distribution of the vector (X_r, Y) as $\mathbb{P}^{X_rY}$ and the marginal distribution of the dimensions X_r and Y as $\mathbb{P}^{X_r}$ and $\mathbb{P}^Y$ respectively, we define

$$\omega_r \stackrel{\text{def}}{=} \gamma_r(\mathbb{P}^{X_rY}, \mathbb{P}^{X_r}\mathbb{P}^Y)$$

to be the measure of dependence between the r^{th} dimension X_r and the response Y. Note that $\gamma_r = 0$ iff X_r is independent of Y and greater the value, greater the degree of dependence. These properties make *sup*-HSIC suitable for ranking the dimensions of X according to the degree of dependence to the response Y. In practice, given n samples, we use the empirical estimator $\widehat{\gamma}$ defined in the previous section. Specifically, we denote the corresponding empirical estimate as $\widehat{\omega}_r = \widehat{\gamma}_r(\mathbb{P}_n^{X_rY}, \mathbb{P}_n^{X_r}\mathbb{P}_n^Y)$.

In order to select the relevant variables (i.e., to estimate $\mathcal{M}$), we first compute $\widehat{\omega}_r$ for $r = 1, \ldots, p_n$ and define

$$\widehat{\mathcal{M}} = \{r : \widehat{\omega}_r \geq cn^{-\kappa}, \text{ for } 1 \leq r \leq p_n\}$$

where $0 \leq \kappa < 1/2$, as the estimated set of relevant features. Note that the set of relevant features is defined as the set of all dimensions that have dependence greater than $cn^{-\kappa}$ with the response. The threshold defined here depends on the value of n. When n is

large, naturally it allows for variables with weaker dependence to be detected.

The approach above has several nice properties. First, the method is model free as it does not assume a specific regression model between X and Y. Second, the response Y may be a vector or more generally a graph or a ranking. As a result, the method can be used for feature selection in the case of multi-label classification and multivariate output regression. Third, the method chooses the kernel k in a principled way by selecting k from a family of positive definite kernels that maximizes the Hilbert Schmidt norm of the covariance operator. Finally, as we show in the next section the method generalizes the recently proposed DC-SIS (Li et al., 2012).

3.1 DC-SIS as a special case of *sup*-HSIC-SIS

In order to see how the proposed method generalizes the recent approach by Li et al. (2012), we appeal to the general equivalence between distance based independence metrics and kernel based independence metrics, as established by Sejdinovic et al. (2012a). To summarize DC-SIS briefly, Li et al. (2012) uses distance covariance metric (Székely et al., 2007) as a measure of independence in the screening approach. In order to see the connection, we first need the following definition due to Lyons (2012).

Definition 1. Let $(\mathcal{X}, \rho_{\mathcal{X}})$ and $(\mathcal{Y}, \rho_{\mathcal{Y}})$ be semi-metric spaces of negative type (cf. Lyons (2012)), with random variables X and Y taking values in $\mathcal{X}$ and $\mathcal{Y}$ respectively. The distance covariance between X and Y is defined as

$$\begin{aligned}\mathrm{dcov}^2(X,Y) = {} & \mathsf{E}_{X,Y}\mathsf{E}_{X',Y'}\rho_{\mathcal{X}}(X,X')\rho_{\mathcal{Y}}(Y,Y') \\ & + \mathsf{E}_X\mathsf{E}_{X'}\rho_{\mathcal{X}}(X,X')\mathsf{E}_Y,\mathsf{E}_{Y'}\rho_{\mathcal{X}}(X,X') \\ & - 2\mathsf{E}_{X,Y}\left(\mathsf{E}_{X'}\rho_{\mathcal{X}}(X,X')\mathsf{E}_{Y'}\rho_{\mathcal{X}}(X,X')\right).\end{aligned}$$

When $\mathcal{X} = \mathbb{R}^s$ and $\mathcal{Y} = \mathbb{R}^t$ with $\rho_{\mathcal{X}} = \rho_{\mathcal{Y}} = \|\cdot - \cdot\|$, dcov reduces to the distance used in (Székely et al., 2007). The following result due to Sejdinovic et al. (2012b) establishes the equivalence between dcov and γ_k.

Theorem 3.1. *Let $(\mathcal{X}, \rho_{\mathcal{X}})$ and $(\mathcal{Y}, \rho_{\mathcal{Y}})$ be semi-metric spaces of negative type with $X \sim \mathbb{P}^X$ and $Y \sim \mathbb{P}^Y$ having joint $\mathbb{P}^{XY}$. Let $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ be kernels on $\mathcal{X}$ and $\mathcal{Y}$ that generate the respective metrics and denote $k((x,y),(x',y')) = k_{\mathcal{X}}(x,x')k_{\mathcal{Y}}(y,y')$. Then $\mathrm{dcov}^2(X,Y) = 4\gamma_k^2(\mathbb{P}^{XY}, \mathbb{P}^X\mathbb{P}^Y)$.*

Example 11 in (Sejdinovic et al., 2012b) shows that $k_q(x,x') = \frac{1}{2}(\|x\|^q + \|x'\|^q - \|x-x'\|^q)$, $x, x' \in \mathbb{R}^d$, $0 < q \leq 2$ generates a semi-metric, $\rho_q(x,x') = \|x-x'\|^q$ of negative type. Choosing $k_{\mathcal{X}} = k_{\mathcal{Y}} = k_1$ yields the *dcov* metric as proposed in (Székely et al., 2007), which is used in DC-SIS. But there is no reason to fix $q = 1$ and it is not possible to know appropriate q a priori, which motivates the use of *sup*-HSIC as a dependence measure in *sup*-HSIC-SIS, thereby generalizing DC-SIS. In addition, the proposed generalization enables one to work with a wide variety of kernel families (and not just $\{k_q : 0 < q \leq 2\}$) and provides a richer set of independence measures between random variables, which in turn enables one to do better model-free feature selection. Thus the proposed *sup*-HSIC-SIS procedure is strictly more general than the DC-SIS method, and achieves better empirical results as demonstrated in Section 6.

4 Theoretical analysis

In this section, we prove the sure screening property of *sup*-HSIC-SIS for $\mathcal{X} \subset \mathbb{R}^{p_n}$ and $\mathcal{Y} \subset \mathbb{R}^d$. Our analysis applies to a range of kernel families and does not impose any moment conditions on the variables X and Y. Further it provides a simpler proof under relaxed assumption compared to Li et al. (2012) even for DC-SIS. For simplicity, we let d to be fixed, but one could also analyze the dependency on d to determine the joint scaling of d and p_n with n. We allow the cardinality of the active set to scale with n, i.e., $|\mathcal{M}_n| = s_n$. The main assumptions we impose are the following:

A1 $\sup\{k_{\mathcal{X}}(x,x) : k_{\mathcal{X}} \in \mathcal{K}_{\mathcal{X}}, x \in \mathcal{X}\} = A < \infty$

A2 $\sup\{k_{\mathcal{Y}}(y,y) : k_{\mathcal{Y}} \in \mathcal{K}_{\mathcal{Y}}, y \in \mathcal{Y}\} = A < \infty$

A3 $\min_{r \in \mathcal{M}} \omega_r \geq 2cn^{-\kappa}$ for some $c > 0$ and $\kappa \in [0, 1/2)$.

Assumption A3 requires that *sup*-HSIC measure corresponding to the relevant variables cannot be too small, which is similar to condition 3 of Fan and Lv (2008) and various other previous works that analyzed marginal screening approaches. The proof of sure screening property of *sup*-HSIC-SIS in Theorem 4.1, uses an intermediate result in Lemma 1 (stated and proved in the appendix).

Definition 2. Let $\mathcal{G}$ be a class of functions on $\mathcal{X} \times \mathcal{X}$ and $\{\rho_1, \ldots \rho_n\}$ be independent Rademacher random variables. The homogeneous Rademacher chaos process of order two with respect to $\{\rho_1, \ldots \rho_n\}$ is defined as $\{n^{-1}\sum_{i<j}^n \rho_i\rho_j g(x_i,x_j) : g \in \mathcal{G}\}$ for some $\{x_1, \ldots, x_n\} \subset \mathcal{X}$. The Rademacher chaos complexity of $\mathcal{G}$ is defined as

$$U_n(\mathcal{G}; \{x_i\}) \stackrel{\mathrm{def}}{=} \mathsf{E}_\rho \sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i<j}^n \rho_i\rho_j g(x_i,x_j) \right|.$$

Theorem 4.1. *Let $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ be measurable kernels satisfying assumptions* **A1** *and* **A2**. *Define $D := \{(x^{(i)}, y^{(i)})\}_{i=1}^n$. Then we have*

$$\begin{aligned}(\mathbb{P}^{XY})^n \left(\left\{ D \in (\mathcal{X} \times \mathcal{Y})^n : \max_{1 \leq r \leq p_n} |\widehat{\omega}_r - \omega_r| \geq cn^{-\kappa} \right) \right\} \\ \leq 6p_n \exp\left(-\frac{\left(cn^{\frac{1}{2}-k} - \mathcal{R}_n - 6A\right)^2}{162A^2} \right), \quad (2)\end{aligned}$$

where $\mathcal{R}_n \stackrel{\text{def}}{=} \sqrt{8AU_n(\mathcal{K}_{\mathcal{Y}}; \{y^{(i)}\})} + \sup_r \left(\sqrt{8U_n(\mathcal{K}; \{(x_r^{(i)}, y^{(i)})\})} + \sqrt{8AU_n(\mathcal{K}_{\mathcal{X}}; \{x_r^{(i)}\})} \right)$. *Furthermore if assumption* **A3** *is also satisfied, then we have the following sure screening property:*

$$(\mathbb{P}^{XY})^n \left(\mathcal{M} \subseteq \widehat{\mathcal{M}} \right) \geq 1 - O\left(s_n e^{-\frac{\left(cn^{\frac{1}{2}-k} - \mathcal{R}_n - 6A \right)^2}{162A^2}} \right).$$

Proof. The proof of (2) follows from applying Lemma 1 from the appendix for each r followed by a union bound. In order to prove the sure screening property, if $\mathcal{M} \nsubseteq \widehat{\mathcal{M}}$, then there must exist some $r \in \mathcal{M}$ such that $\widehat{\omega}_r < cn^{-\kappa}$. But, from the assumption **A3**, we have that $|\widehat{\omega}_r - \omega_r| > cn^{-\kappa}$ for some $r \in \mathcal{M}$. Hence we note that the event $\{\mathcal{M} \nsubseteq \widehat{\mathcal{M}}\}$ happens if $\{|\widehat{\omega}_r - \omega_r| > cn^{-\kappa}\}$, for some $r \in \mathcal{M}$. Define Γ to be the event $\{\max_{r \in \mathcal{M}} |\widehat{\omega}_r - \omega_r| \leq cn^{-\kappa}\}$. Then we have $(\mathbb{P}^{XY})^n \left(\mathcal{M} \subseteq \widehat{\mathcal{M}} \right) \geq (\mathbb{P}^{XY})^n(\Gamma)$ and the following sequence of inequality holds. Define $\Pr \stackrel{\text{def}}{=} (\mathbb{P}^{XY})^n$.

$$\begin{aligned}
\Pr(\Gamma) &= 1 - \Pr(\Gamma^c) = 1 - \Pr\left(\min_{r \in \mathcal{M}} |\widehat{\omega}_r - \omega_r| \geq cn^{-\kappa} \right) \\
&= 1 - s_n \Pr\left(|\widehat{\omega}_r - \omega_r| \geq cn^{-\kappa} \right) \\
&\geq 1 - O\left(s_n \exp\left(-\frac{\left(cn^{-\kappa+1/2} - \mathcal{R}_n - 6A \right)^2}{162A^2} \right) \right).
\end{aligned}$$

This completes the proof. □

Note that an important quantity controlling the rates is the term $\mathcal{R}_n$ that involves the Rademacher chaos complexities of $\mathcal{K}$, $\mathcal{K}_{\mathcal{X}}$ and $\mathcal{K}_{\mathcal{Y}}$. Sriperumbudur et al. (2009) has shown that for VC-subgraph classes of kernels, the Rademacher chaos complexity is bounded above by a constant that depends on the VC dimension of the class. Examples of such kernel classes in a d-dimensional Euclidean space include Gaussian, Laplacian, Matern class etc. We refer the reader to (Sriperumbudur et al., 2009) for a detailed discussion and several more examples. In our setting, if $\mathcal{K}$, $\mathcal{K}_{\mathcal{X}}$ and $\mathcal{K}_{\mathcal{Y}}$ are VC subgraph classes, then $\Pr\left(\max_{1 \leq r \leq p_n} |\widehat{\omega}_r - \omega_r| \geq cn^{-\kappa}\right) \leq O\left(p_n \exp(-c_1 n^{1-2\kappa})\right)$ from which we observe that the proposed approach enables us to handle ultrahigh dimensionality, i.e., $\log p_n = o(n^{1-2\kappa})$.

In order to control the false positive rates, if we assume that $\max_{r \notin \mathcal{M}} |\omega_r| = O(n^{-\kappa})$, then with probability tending to 1, we have

$$\max_{r \notin \mathcal{M}} |\widehat{\omega}_r| \leq C(n^{-\kappa}).$$

for some constant $C > 0$. By applying Theorem 4.1, we have: $\Pr(\mathcal{M} = \widehat{\mathcal{M}}) = 1 - O(1)$. This gives a model selection consistency result under the assumption that there is a strict separation between the set of relevant and irrelevant variables. But to be more general, we analyze below, the cardinality of the set $\widehat{\mathcal{M}}$.

4.1 Upper bounding the cardinality of $\widehat{\mathcal{M}}$

A main reason for performing feature screening is to reduce the dimensionality from exponential to something that could be handled, say polynomial with the sample size. With that one could use cleaning procedures to further refine the feature selection process. In this section, we show that by appropriately selecting the bound on the kernel (i.e., the value A), one could make the cardinality of the estimated set grow polynomially in the sample size. Specifically, we have the following theorem.

Theorem 4.2. *Let $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ be measurable kernels satisfying assumptions* **A1** *and* **A2**. *Then there exists a constant $c > 0$ such that,*

$$(\mathbb{P}^{XY})^n \left(|\widehat{\mathcal{M}}| \leq O(n^{\kappa} p_n A) \right) \geq 1 - p_n e^{-\frac{\left(cn^{\frac{1}{2}-k} - \mathcal{R}_n - 6A \right)^2}{162A^2}}.$$

Proof. First we note that $\sum_{r=1}^{p_n} \omega_r \leq p_n \max_r \omega_r \leq CAp_n = O(Ap_n)$. Now this would imply that $|\{r : \omega_r > \epsilon n^{-\kappa}\}|$ cannot exceed $O(n^{\kappa} Ap_n)$ for any $\epsilon > 0$. Thus on the set, $\Upsilon = \{\max_{1 \leq r \leq d} |\widehat{\omega}_r - \omega_r| \leq \epsilon n^{-\kappa}\}$, $|\{r : \widehat{\omega}_r > 2\epsilon n^{-\kappa}\}|$ cannot exceed $|\{r : \omega_r > \epsilon n^{-\kappa}\}|$, which would be bounded by $O(n^{\kappa} Ap_n)$. If we take $\epsilon = c/2$, we have $\Pr(|\widehat{\mathcal{M}}| \leq O(n^{\kappa} Ap)) \geq \Pr(\Upsilon)$ and the conclusion follows from (2). □

The main consequence of the above theorem is that when $A = O(n^{\tau}/p_n)$, for some $\tau > 0$, then we have $|\widehat{\mathcal{M}}| = O(n^{\kappa+\tau})$ and thus the size of the selected set is of polynomial order in n. Compared to the initial case when the dimensionality is of exponential order, this is a huge improvement in terms of feature selection. This also gives us some insights on how to design or select kernels such that we could have a control over the cardinality of the selected feature set size.

5 Iterative Screening procedures

Any screening method based on marginal computations suffers from the following problems (cf. (Fan et al., 2009)): (1) any irrelevant covariate that is highly correlated with the set of relevant covariates could be selected and (2) marginally uncorrelated covariate that is jointly correlated with the response might not be selected. Here, we propose two approaches that could be used in order to handle such scenarios.

5.1 Method 1

We first consider the situation when important covariates are marginally weakly correlated, but jointly correlated to the response. In order to deal with this situation, we propose the following iterative method:

1. Compute *sup*-HSIC between each dimension and response and select the covariates that have $\omega_r > \lambda_t$. Let $\widehat{\mathcal{M}}_{(t)}$ be the set of selected covariates at round t with $X_{\widehat{\mathcal{M}}_{(t)}}$ being the set of selected features.
2. Compute *sup*-HSIC between $(Y, (X_{\widehat{\mathcal{M}}_{(t)}}, X_j))$ and marginal Y and $(X_{\widehat{\mathcal{M}}_{(t)}}, X_j))$ for all $j \in \widehat{\mathcal{M}}^c_{(t)}$. The selected feature set $\widehat{\mathcal{M}}'_{(t)}$ consists of covariates j for which the above calculated *sup*-HSIC is greater than the *sup*-HSIC between $(Y, X_{\widehat{\mathcal{M}}_{(t)}})$ and the marginal Y and $X_{\widehat{\mathcal{M}}_{(t)}}$. Update $\widehat{\mathcal{M}}_{(t)} = \widehat{\mathcal{M}}_{(t-1)} \cup \widehat{\mathcal{M}}'_{(t)}$
3. Repeat the procedure till $\widehat{\mathcal{M}}_{(t)} = \widehat{\mathcal{M}}_{(t-1)}$ or until $|\bigcup_t \widehat{\mathcal{M}}_{(t)}| > n$.

In the above iterative approach, the threshold λ_t is set at a high value during the initial rounds and reduced as the rounds progress. In practice, it could be selected using cross-validation. Heuristics for selecting the threshold for such iterative methods could be found in (Fan et al., 2009). The above iterative approach would be able to detect covariates that are marginally uncorrelated with the response (and hence not selected in initial rounds), but are jointly correlated because we measure *sup*-HSIC between the joint vector $(X_{\widehat{\mathcal{M}}_{(t)}}, X_j)$ and the response Y.

5.2 Method 2

This approach is motivated by the iterative screening procedure proposed by Fan and Lv (2008) which was based on residuals computed between the covariates and response under a linear model assumption. It is not possible to directly adopt such a procedure in our case, as the proposed approach is model-free. In order to proceed, first we introduce the input residual matrix. Let $\mathbf{X}_{\widehat{\mathcal{M}}_{(t)}} \in \mathbb{R}^{n \times |\widehat{\mathcal{M}}_{(t)}|}$ be data matrix associated with selected covariates at round t and $\mathbf{X}_{\widehat{\mathcal{M}}^c_{(t)}} \in \mathbb{R}^{n \times (p - |\widehat{\mathcal{M}}_{(t)}|)}$ be data matrix corresponding to the remaining covariates. The input residual matrix is defined as the projection of complement of selected variables in a particular step onto the orthogonal complement space of the selected variables in that step, i.e., $\mathbf{X}^{(t)}_r = \{\mathbf{I}_{n\times n} - \mathbf{X}_{\widehat{\mathcal{M}}_{(t)}}(\mathbf{X}^\top_{\widehat{\mathcal{M}}_{(t)}}\mathbf{X}_{\widehat{\mathcal{M}}_{(t)}})^{-1}\mathbf{X}^\top_{\widehat{\mathcal{M}}_{(t)}}\}\mathbf{X}_{\widehat{\mathcal{M}}^c_{(t)}}$. The key idea of this approach is that the input residual matrix at a particular step is uncorrelated with the space of selected variables in that step. Thus covariates that would have been selected because they are correlated with a true relevant covariate (and hence correlated with the response) could be avoided in this approach. This discussion leads to the following approach:

1. Calculate *sup*-HSIC to the original data set and let $\widehat{\mathcal{M}}_{(t)}$ be set of selected features at round t.
2. Compute the residual data matrix, $\mathbf{X}^{(t)}_r = \{\mathbf{I}_{n\times n} - \mathbf{X}_{\widehat{\mathcal{M}}_{(t)}}(\mathbf{X}^\top_{\widehat{\mathcal{M}}_{(t)}}\mathbf{X}_{\widehat{\mathcal{M}}_{(t)}})^{-1}\mathbf{X}^\top_{\widehat{\mathcal{M}}_{(t)}}\}\mathbf{X}_{\widehat{\mathcal{M}}^c_{(t)}}$ and compute *sup*-HSIC between $\mathbf{X}^{(t)}_r$ and the response to obtain the selected feature set $\widehat{\mathcal{M}}'_{(t)}$ and update $\widehat{\mathcal{M}}_{(t)} = \widehat{\mathcal{M}}_{(t-1)} \cup \widehat{\mathcal{M}}'_{(t)}$. Stop when $\widehat{\mathcal{M}}_{(t)} = \widehat{\mathcal{M}}_{(t-1)}$ or $|\bigcup_t \widehat{\mathcal{M}}_{(t)}| > n$.

Similar to Method 1, the threshold for the initial rounds are set at high value and subsequently lowered. Since the residual matrix at each step is not correlated with the selected covariates, the covariates that are strongly correlated with any of true active covariates would not be selected. Also covariates that were actually correlated to the response (but were not selected) would now be detected easily.

6 Experiments

In this section, we report experimental results on various synthetic and real-world data sets to demonstrate the advantage of the proposed approach (*sup*-HSIC-SIS) over various feature screening approaches. For the experiments on synthetic data, we consider data settings from (Li et al., 2012) in order to make a direct comparison to their approach. For evaluation on real-world data, we consider a very high dimensional gene data set and a multi-label data set and show that the proposed approach performs significantly better than the existing approaches.

6.1 Synthetic data – univariate response

The synthetic data set is generated as follows: $X \sim N(0, \Sigma)$ where $\Sigma \in \mathbb{R}^{p\times p}$ with entries $\sigma_{i,j} = 0.8^{|i-j|}$. We set $n = 200$ and let p to be 5000. We generate the response Y according to three models:

1. $Y = c_1\beta_1 X_1 X_2 + c_3\beta_2 1(X_{12} < 0) + c_4\beta_3 X_{22} + \epsilon$
2. $Y = c_1\beta_1 X_1 X_2 + c_3\beta_2 1(X_{12} < 0) X_{22} + \epsilon$
3. $Y = c_1\beta_1 X_1 + c_2\beta_2 X_2 + c_3\beta_3 1(X_{12} < 0) + \exp(c_4|X_2 2|)\epsilon$

where $\beta_j = (-1)^U(a + |Z|)$ where $a = 4\log n/\sqrt{n}$, $U \sim$ Bernoulli(0.4) and $Z, \epsilon \sim N(0.1)$. Note that all models are non-linear in X_{12}. Further the third model is heteroscedastic. For the fourth data set, the relationship between the response and covariates is given by the following joint model, $\mathbb{P}^{X_r Y} \propto 1 + \sin(lx)\sin(ly)$ for integer l, on the support $[-\pi, \pi] \times [-\pi, \pi]$ for each r. Note that when $l = 0$, X_r and Y are independent

NIS	$q=1$	$q=\frac{1}{2}$	$q=\frac{1}{4}$	$\sup_q k_q$	Gauss.
$P(\mathcal{M}^*=\widehat{\mathcal{M}})$					
0.78	0.79	0.82	0.84	0.88	0.87
0.73	0.75	0.79	0.80	0.83	0.84
0.73	0.73	0.75	0.78	0.82	0.82
0.35	0.40	0.52	0.60	0.71	0.80
$P(\mathcal{M}^*\subset\widehat{\mathcal{M}})$					
0.96	0.98	1.00	1.00	1.00	1.00
0.94	0.95	0.99	1.00	1.00	1.00
0.93	0.96	1.00	1.00	1.00	1.00
0.6	0.69	0 .72	0.75	0.92	0.98
$\lvert\widehat{\mathcal{M}}\rvert$					
10.1	7.4	5.4	4.4	4.2	4.2

Table 1: Probability of support recovery using the distance kernel and Gaussian kernel: First four rows correspond to $P(\mathcal{M}^*=\widehat{\mathcal{M}})$ (corresponding to models 1, 2, 3 and 4 respectively) and the last four rows correspond to $P(\mathcal{M}^*\subset\widehat{\mathcal{M}})$. The very last row corresponds to the average cardinality of selected set.

and as $|l|$ increases they become dependent wherein the joint distribution departs from the uniform at higher frequencies, making it hard to detect from small sample sizes. We set $l=10$ for $r=1,2,3,4$ and $l=0$ for the rest. This way the response is dependent on the first four covariates only.

We compared the following approaches: HSIC-SIS with $k_q(z,z')=1/2\left(\|z\|^q+\|z'\|^q-\|z-z'\|^q\right)$ at $q=1,0.5,0.25$, *sup*-HSIC-SIS with $\mathcal{K}=\{k_q:0<q\leq 2\}$, *sup*-HSIC-SIS with a Gaussian kernel and nonparametric independence screening (NIS) of Fan et al. (2011). Note that $q=1$ corresponds to DC-SIS. Table 1 shows $P(\mathcal{M}^*\subset\widehat{\mathcal{M}})$ and $P(\mathcal{M}^*=\widehat{\mathcal{M}})$ computed over 500 experiments. Note that the proposed *sup*-HSIC-SIS approach performs better than other approaches. Also in some situations the Gaussian kernel performs better, while in some the distance kernel performs better. Further, the advantage of the proposed approach is clearly demonstrated in the fourth model, where the other approaches are not able to detect the specific type of dependency whereas the proposed approach with Gaussian kernel performs the best. Selecting a kernel for a given task is a more involved problem which we hope to address in the future (a simple step in this direction would be to consider a convex combination of base kernels).

6.2 Synthetic data – multivariate response

In this experiment, we deal with multivariate outputs, while we generate X as before. We generate Y from normal distribution with mean zero and conditional covariance matrix $\Sigma_{Y|X}$ given by $\sigma_{11}=\sigma_{22}=1$ and $\sigma_{12}=\sigma_{21}=\sigma(X)$. We consider two correlation functions for $\sigma(X)$ given by

1. $\sigma(X)=\sin(\beta_1^\top X)$ where $\beta_1=(0.8,0.6,0,\ldots,0)$
2. $\sigma(X)=\{\exp(\beta_2^\top X-1)/\exp(\beta_2^\top X+1)\}$ where $\beta_2=(2-U_1,2-U_2,2-U_3,2-U_4,0,\ldots,0)$ with U_i drawn i.i.d. from Uniform$[0,1]$.

$q=1$	$q=\frac{1}{2}$	$q=\frac{1}{4}$	$\sup_q k_q$	Gaussian
$P(\mathcal{M}^*=\widehat{\mathcal{M}})$				
0.79	0.85	0.86	0.91	0.90
0.77	0.81	0.85	0.87	0.89
$P(\mathcal{M}^*\subset\widehat{\mathcal{M}})$				
0.97	0.99	1.00	1.00	1.00
0.96	0.97	0.98	1.00	1.00
$\lvert\widehat{\mathcal{M}}\rvert$				
9.4	6.7	5.2	4.3	4.4

Table 2: Probability of support recovery using the distance kernel and Gaussian kernel. First two rows correspond to $P(\mathcal{M}^*=\widehat{\mathcal{M}})$ and the last three rows correspond to $P(\mathcal{M}^*\subset\widehat{\mathcal{M}})$. The very last row corresponds to the average cardinality of selected set over all experiments.

Note that for this experiment, the NIS method could not be used directly as it cannot handle multivariate outputs. Hence, we only compared our approach to DC-SIS, whose results are presented in Table 2. It is clear from Table 2 that *sup*-HSIC-SIS performs better in this setup as well.

6.3 Synthetic data – Iterative screening

In this section, we demonstrate the advantage of the iterative screening procedures (see Section 5) over *sup*-HSIC-SIS, using a Gaussian kernel. We use the simulation setup provided by Fan and Lv (2008) which consists of a linear model $y=\beta^\top x+\epsilon$ with $\beta\in\mathbb{R}^p$ and $\epsilon\sim N(0,1)$. We set $\beta=(5,5,5,-15\sqrt{\rho},0,\ldots,0)$ with $p=2000$ and we draw $n=100$ covariates x from a mean zero normal distribution with $\Sigma_{p\times p}=\sigma_{ij}$, with entries $\sigma_{ii}=1$ for $i=1,\ldots,p$ and $\sigma_{i4}=\sigma_{4i}=\sqrt{\rho}$ for $i\neq 4$ and $\sigma_{ij}=\rho$ for $i\neq j$, $i\neq 4$ and $j\neq 4$. Note that all predictors except x_4 are equally correlated with correlation coefficient ρ. In addition, x_4 has correlation coefficient ρ with all other predictors and is independent of y, but x_4 belongs to the active set when $\rho\neq 0$. We vary ρ to be $0,0.1,0.5,0.9$.

We perform 2 iterations of both the iterative algorithms as we attain the stopping criterion. The threshold parameter was set based on cross-validation. We repeat the experiment for 1000 trials and report the probability of including all correct variables in the estimated set ($P(\mathcal{M}^*\subset\widehat{\mathcal{M}})$) (see Table 3). Note that the non-iterative version performs poorly. In fact it could select all active covariates only by chance. Both method 1 and 2 perform well in this situation, as ex-

ρ	0	0.1	0.5	0.9
sup-HSIC-SIS	0.98	0.89	0.54	0.42
Method 1	1.00	1.00	0.99	0.95
method 2	1.00	1.00	1.00	1.00

Table 3: Advantage of iterative methods over sup-HSIC-SIS. The values reported are estimates of $P(\mathcal{M}^* \subset \widehat{\mathcal{M}})$ over 1000 trials.

pected. Method 1 performs slightly worse compared to Method 2 because it has to deal with multivariate *sup*-HSIC evaluations in the second step, which is comparatively hard with less samples.

6.4 Gene array data set

Furthermore we analyze the Affymetric GeneChip Rat Genome 230 2.0 Array data set which was previously used by Scheetz et al. (2006) and Huang et al. (2010). This data set consists of 120 rat subjects from which $18,975$ different probes sets (genes) from eye tissue were measured. Following Huang et al. (2010), the intensity values were normalized and gene expression levels were analyzed on a logarithmic scale. Specifically, we are interested in finding the genes that are most related to TRIM32 gene, the reason being that this gene was recently found to cause Bardet-Biedl syndrome, a topic of interest in the biological community. The data set is highly challenging with $n = 120$ and $p = 18,975$ with non-linear relationships.

We used *sup*-HSIC-SIS with Gaussian kernel to select the important genes and compared it to BA-HSIC, NIS and DC-SIS methods. BA-HSIC cannot actually handle high dimensionality because of its design; we just use it for comparison purpose. For the experiment, we used 100 training samples to select the features (genes), and fitted an additive model (with functions in Sobolev classes) using the selected features, and compared the predictive error (PE) on the remaining 20 points. BA-HSIC performs poorly in the regime considered (small n, large p) and fails to select many important genes (that are selected by all the other methods) in addition to exhibiting a relatively poor predictive error. Both NIS and DC-SIS select 8 genes, whereas the proposed approach selects 7 genes. Also, the predictive accuracy of the proposed approach is smaller implying that maybe the additional gene selected by the other methods is not actually necessary to explain the response considered. Thus the proposed approach would present a biologist to work with a more targeted set of genes for subsequent investigations.

6.5 Multi-label classification data set

For the next experiment, we choose to evaluate the performance of *sup*-HSIC-SIS (using Gaussian kernel), DC-SIS and BA-HSIC on 4 different yahoo multi-label data sets: arts, business, education and health (Ueda and Saito, 2003). The task is to select features first using the above three methods and perform classification in the next step using one-vs-all multi-label SVM approach. For each of the data sets, the number of samples was set at $n = 1000$ during training stage. The samples were selected such that the class labels were balanced. The dimensionality of (X, Y) for the data sets are $(17973, 19), (16621, 17), (20782, 14), (18430, 14)$ respectively. Table 5 shows the classification accuracy and the cardinality of the selected features for different data sets. Note that the proposed approach achieves better classification accuracy with lesser number of features demonstrating the wide applicability of the proposed approach.

Method	Cardinality	PE
BA-HSIC	12.32	4.32
NIS	7.73	0.47
DC-SIS	7.21	0.45
sup-HSIC-SIS	6.76	0.39

Table 4: Gene data set: Cardinality of selected set and predictive error (PE) under an additive model.

Data set	BA-HSIC	DC-SIS	Proposed
Arts	(967) 25.87	(658) 14.32	(435) 9.54
Business	(1231) 26.32	(743) 15.64	(611) 10.11
Edu	(1123) 21.02	(643) 11.31	(533) 9.21
Health	(1045) 22.54	(764) 13.42	(564) 10.74

Table 5: Test set classification error on the multi-label data sets. The number in the bracket correspond to the cardinality of selected feature set.

7 Discussion

We proposed an RKHS embedding approach for feature screening of ultrahigh dimensional data. The proposed approach, which is a strict generalization of the procedure recently proposed in (Li et al., 2012), is model-free and works with multivariate and non-standard output spaces (like graphs or rankings). We proved the feature screening consistency of the proposed approach and empirically demonstrated its capability in handling ultrahigh dimensional regimes on various synthetic and real-world data sets. Furthermore, we proposed two iterative screening methods to counter some problems exhibited by the marginal screening based feature selection approaches.

Future work includes a theoretical analysis of the proposed iterative procedures and to develop other iterative screening procedures that would also enable one to exclude already selected features/covariates in future.

References

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, London, UK.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351.

Fan, J., Feng, Y., and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *J. Amer. Statist. Assoc.*, 106(494):544–557.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. Roy. Statist. Soc. Ser. B*, 70:849–911.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20:101–148.

Fan, J., Samworth, R., and Wu, Y. (2009). Ultrahigh dimensional feature selection: Beyond the linear model. *J. of Machine Learning Research*, 10:2013–2038.

Gretton, A., Borgwardt, K. M., Rasch, M., Schölkopf, B., and Smola, A. (2007). A kernel method for the two sample problem. In *Advances in Neural Information Processing Systems 19*, pages 513–520.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In *Proc. of the 16th International Conference on Algorithmic Learning Theory*, pages 63–77.

Huang, J., Horowitz, J., and Wei, F. (2010). Variable selection in nonparametric additive models. *Annals of statistics*, 38(4):2282–2313.

Ji, P. and Jin, J. (2012). UPS delivers optimal phase diagram in high-dimensional variable selection. *Annals of Statistics*, 40(1):73–103.

Li, R., Zhong, W., and Zhu, L. (2012). Feature screening via distance correlation learning. *J. Amer. Statist. Assoc.*, 107:1129–1139.

Lyons, R. (2012). Distance covariance in metric spaces. *Annals of Probability*. To appear.

Nishiyama, Y., Boularias, A., Gretton, A., and Fukumizu, K. (2012). Hilbert space embeddings of POMDPs. In *Proc. 28th Conference on Uncertainty in Artificial Intelligence*, pages 644–653.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). Sparse additive models. *J. Roy. Statist. Soc. Ser. B*, 71:1009–1030.

Scheetz, T., Kim, K., Swiderski, R., Philp, A., Braun, T., Knudtson, K., Dorrance, A., DiBona, G., Huang, J., Casavant, T., Sheffield, V., and Stone, E. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences*, 103(39):14429–14434.

Sejdinovic, D., Gretton, A., Sriperumbudur, B., and Fukumizu, K. (2012a). Hypothesis testing using pairwise distances and associated kernels. pages 1111–1118.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2012b). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *arXiv:1207.6076*.

Smola, A. J., Gretton, A., Song, L., and Schölkopf, B. (2007). A Hilbert space embedding for distributions. In *Proc. 18th International Conference on Algorithmic Learning Theory*, pages 13–31. Springer-Verlag, Berlin, Germany.

Song, L., Smola, A., Gretton, A., Bedo, J., and Borgwardt, K. (2012). Feature selection via dependence maximization. *J. of Machine Learning Research*, 13:1393–1434.

Sriperumbudur, B., Fukumizu, K., Gretton, A., Lanckriet, G., and Schölkopf, B. (2009). Kernel choice and classifiability for RKHS embeddings of probability distributions. In *Advances in Neural Information Processing Systems 22*, pages 1750–1758. MIT Press.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. G. (2010). Hilbert space embeddings and metrics on probability measures. *J. of Machine Learning Research*, 11:1517–1561.

Székely, G., Rizzo, M., and Bakirov, N. (2007). Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35(6):2769–2794.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58:267–288.

Ueda, N. and Saito, K. (2003). Parametric mixture models for multi-labeled text. In *Advances in Neural Information Processing Systems 15*, pages 721–728.

Zhao, P. and Yu, B. (2007). On model selection consistency of lasso. *J. of Machine Learning Research*, 7:2541–2563.